ED 418 101 TM 028 195

AUTHOR	Chung, Gregory K. W. K.; O'Neil, Harold F., Jr.
TITLE	Methodological Approaches to Online Scoring of Essays.
INSTITUTION	National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-461
PUB DATE	1997-12-00
NOTE	39p.
CONTRACT	R305B60002-97
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Automation; *Computer Assisted Testing; *Essays; *Scoring; Semantics; *Standardized Tests; Test Scoring Machines; *Writing Tests

ABSTRACT

This report examines the feasibility of scoring essays using computer-based techniques. Essays have been incorporated into many of the standardized testing programs. Issues of validity and reliability must be addressed to deploy automated approaches to scoring fully. Two approaches that have been used to classify documents, surface- and word-based techniques, are reviewed. The candidate approaches to the automated classification of documents are reviewed, and then how these approaches could be used to achieve the overarching goal of the automated scoring of essays is discussed. The two approaches considered are Project Essay Grade (PEG) (A. Daigon, 1966; E. Page, 1966, 1968, 1994; E. Page and N. Peterson, 1995) and latent semantic analysis (LSA) (P. Foltz, 1996; T. Landauer, D. Laham, B. Rehder, and M. Schreiner, 1997). PEG uses a regression model where the independent variables are surface features of the text (document length, word length, and punctuation) and the dependent variable is the essay score, and LSA is based on a factor-analytic model of word co-occurrences. Following the review of PEG and LSA, additional uses of automated scoring of text-based data are explored. The final section outlines a plan for a feasibility study of the automated processing of text-based data. One assumption of this report and of both scoring approaches is that the human rating is the best estimate of the true essay score, and that there will almost always be a need for some portion of the documents to be scored using multiple trained raters. The potential of the proposed investigation lies more in the potential for practical spin-offs than in any theoretical contribution to writing, education, assessment, or cognition. (Contains 5 figures, 2 tables, and 50 references.) (SLD)

CRESST

National Center for Research
on Evaluation, Standards,
and Student Testing

Methodological Approaches to Online Scoring of Essays

CSE Technical Report 461

Gregory K. W. K. Chung
CRESST/University of California, Los Angeles

Harold F. O'Neil, Jr.
University of Southern California/CRESST

▸ UCLA Center for the Study of Evaluation

in collaboration with:

▸ University of Colorado
▸ NORC, University of Chicago
▸ LRDC, University of Pittsburgh
▸ University of California, Santa Barbara
▸ University of Southern California
▸ The RAND Corporation

Methodological Approaches to Online Scoring of Essays

CSE Technical Report 461

Gregory K. W. K. Chung
CRESST/University of California, Los Angeles

Harold F. O'Neil, Jr.
University of Southern California/CRESST

December 1997

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

METHODOLOGICAL APPROACHES TO ONLINE SCORING OF ESSAYS

Gregory K. W. K. Chung
CRESST/University of California, Los Angeles

Harold F. O'Neil, Jr.
University of Southern California/CRESST

In this report, we examine the feasibility of scoring essays using computer-based techniques. We review two approaches that have been used to classify documents, surface- and word-based analysis techniques. We omit other text analysis techniques, such as content analyses (Krippendorff, 1980; Roberts & Popping, 1993), neural networks (Chen, 1995; Chen, Orwig, Hoopes, & Nunamaker, 1994; Lin, Soergel, & Marchionini, 1991; Orwig, Chen, & Nunamaker, 1997; Ritter & Kohonen, 1989), and artificial intelligence approaches that attempt to understand the meaning of the text (Carley, 1988; Dreyfus, 1992; Kaplan & Bennett, 1994; Martinez & Bennett, 1992). Our focus is on methods that perform text classification rather than text interpretation. The objective of this report is to first review candidate approaches to the automated classification of documents, and then outline how these approaches could be used to achieve our overarching goal of the automated scoring of essays. Our stance is to evaluate these approaches from an applied perspective. We evaluate strengths and weaknesses of each approach and consider both near- and long-term issues of availability, application to text processing in general, and computational complexity.

In the remainder of this section we briefly identify the problem with scoring essays, specify requirements for an automated scoring system, and make explicit our assumptions about such a system. Next, we review two existing approaches to electronic scoring of essays, Project Essay Grade (PEG) (Daigon, 1966; Page, 1966, 1968, 1994; Page & Peterson, 1995) and latent semantic analysis (LSA) (Foltz, 1996; Landauer, Laham, Rehder, & Schreiner, 1997). Both methods were developed along independent research lines. Briefly, PEG uses a regression model where the independent variables are surface features of the text (e.g., document length, word length, punctuation) and the dependent variable is the essay score; LSA is based on a factor-analytic model of word co-occurrences. Following the review of PEG and LSA, we speculate on additional uses of automated scoring of text-based data. In the final section we outline a plan for a feasibility study.

The Problem With Scoring Essays

Assessing student performance using written responses has many desirable characteristics, particularly in that students are required to construct a response. This assessment method requires more on the part of students than multiple-choice or selection tasks. Students have to generate a response that demonstrates their understanding and synthesis of the content, and their use of analytic and logical thinking skills to develop a coherent argument. However, despite these benefits, the scoring of essays is costly in terms of time and resources. Grading essays is labor intensive and time consuming. For high-stakes assessments, each essay is scored at least twice by independent raters. Multiple raters have to be recruited, trained and calibrated. Hardy (1995) estimates that the cost of scoring essays ranges from $3 to $6 per essay, using a holistic rubric at a rate of 12 minutes per essay. Other factors that increase scoring costs include number of raters per essay, length of essay, complexity of student response, and whether analytical scoring and diagnostic reporting are done. Because of the effort involved in scoring essays, there is a substantial lag between test administration and test reporting. What is needed is a system that (a) preserves the benefits of students constructing written responses, (b) can predict essay scores comparable to human raters, (c) increases essay scoring throughput, and (d) reduces the overall cost of scoring essays. We next elaborate on some of these requirements.

System Requirements

In this section we lay out three requirements we expect of an automated scoring system. First and foremost, the scoring system must accurately classify documents. This is the most obvious yet most important requirement. By accurate, we mean the system should assign the same score to the essay as would a trained rater. Without this requirement being satisfied, the scores predicted by the system cannot be referenced to a known performance standard.

A second requirement is that the system should accurately classify similar documents as belonging to the same group. This requirement is independent of the accuracy of classifying a document relative to a human rater. This requirement specifies that the system should be able to compare two documents and compute a measure of their similarity.

Finally, the system should have well-behaved system parameters. For example, alternating between different term-weighting methods should result in repeatable classification performance. The effects of variation between different parameters should follow predictable curves. The specific parameters will be discussed later.

Design Assumptions

The assumptions behind these requirements are based on the constraint that the scoring system will not attempt to understand the text; rather, the scoring system will attempt to classify text relative to other text. We discuss our assumptions next.

Human ratings are the best estimate of the true score. This assumption is at the foundation of the scoring system. Because our focus is on classification, there needs to be a set of documents to be referenced against. We assume human-scored documents to be the standard against which to evaluate the performance of the scoring system. Thus, correlations between the predicted scores and the human scores are used as the performance measure.

Note that this assumption is short-term. We acknowledge Bennett and Bejar's (1997) caution not to rely exclusively on human scores as the criterion for judging system performance. However, at this time we have no other criterion against which to compare automated scoring performance.

The unscored documents reflect a systematic response. This assumption means that the documents to be processed reflect a reasonable attempt at answering the essay prompt. This aspect is critical, because the PEG and LSA systems ignore linguistic factors to varying degrees.

Focus on text classification, not text interpretation. The goal of the system is document classification, not document understanding. The system attempts to identify other similar documents (along some facet), but no attempt is made to understand the meaning of the document itself. Thus, while the system would be able to generate a similarity score between one document and another, no knowledge base is maintained. This assumption is the basis for excluding artificial intelligence approaches, which have yet to result in general solutions to understanding natural language (Dreyfus, 1992; Kaplan & Bennett, 1994).

In the next section we examine two approaches to essay scoring, PEG and LSA. We present results of their efforts, focusing on the correlations between the scores predicted by the system and the scores assigned by human raters. We also point out strengths and weaknesses of each approach.

Review of Existing Approaches of Automated Scoring of Essays

We have identified two methodologies arising from independent efforts in education and information science. The first method, developed by Ellis Page over 30 years ago, uses measures derived from the surface features of an essay. The second approach, latent semantic analysis, uses word co-occurrences as the basis for comparing two bodies of text.

Project Essay Grade (PEG)

Project Essay Grade was developed in the early 1960s by Ellis Page (Daigon, 1966; Page, 1966, 1968, 1994; Page & Petersen, 1995). Page's method of scoring essays is to use what Page calls "proxy" measures as an estimate of some intrinsic quality of writing style. Proxies are used because of the computational difficulty in obtaining direct measures of writing style. PEG assumes proxies reflect use of a particular writing construct. For example, diction (defined as appropriate word choice) cannot be measured directly by PEG. Instead, PEG measures the proportion of uncommon words in an essay and uses this measure as a proxy or estimate of diction. As the software gets more sophisticated, presumably more direct measures of writing style can be incorporated.

Table 1 lists the proxy variables used in the initial PEG system (Page, 1968). Also listed are the correlations between the proxy variable and the human-rated essay score, and the beta weights associated with the regression equation. Subsequent work (Page, 1994; Page & Peterson, 1995) has refined the list of proxy variables, but little information on the specific variables has been published. Interestingly, the measures most positively associated with essay scores were standard deviation of word length (*r* = .53) followed by the average word length (*r* = .51), number of commas (*r* = .34), essay length in words (*r* = .32), number of prepositions (*r* = .25), and number of dashes (*r* = .22). The proxies negatively associated with essay score are the number of uncommon words

Table 1

List of Proxy Variables for the First PEG System (Page, 1968, p. 216)

Proxy variables	Correlation with essay score (human-rated)	Beta weights
Title present	0.04	0.09
Average sentence length	0.04	-0.13
Number of paragraphs	0.06	-0.11
Subject-verb openings	-0.16	-0.01
Length of essay in words	0.32	0.32
Number of:		
Parentheses	0.04	-0.01
Apostrophes	-0.23	-0.06
Commas	0.34	0.09
Periods	-0.05	-0.05
Underlined words	0.01	0.00
Dashes	0.22	0.10
Colons	0.02	-0.03
Semicolons	0.08	0.06
Quotation marks	0.11	0.04
Exclamation marks	-0.05	0.09
Question marks	-0.14	0.01
Prepositions	0.25	0.10
Connectives	0.18	-0.02
Spelling errors	-0.21	-0.13
Relative pronouns	0.11	0.11
Subordinating conjunctions	-0.12	0.06
Common words on Dale[a]	-0.48	-0.07
Sentences' end punctuation present	-0.01	-0.08
Hyphens	0.18	0.07
Slashes	-0.07	-0.02
Average word length in letters	0.51	0.12
Standard deviation of word length	0.53	0.30
Standard deviation of sentence length	-0.07	0.03

Note. From Page, 1968, p. 216.

[a] Dale's list contains the 1,000 most common words used in the English language.

(r = –.48), followed by number of apostrophes (r = –.23) and number of spelling errors (r = –.21). Page (1968) did not report whether any of these variables were significant.

Essay Scoring With PEG

Scoring essays with the PEG system begins with first scoring essays using trained raters. The actual essays are then converted to electronic form and all proxies measured for each essay. The data are then entered into a multiple regression analysis, with the independent variables being the proxy measures and the dependent variable being the human-rated essay score. The weights derived in this stage are used in the next step. For the remaining essays, all proxy variables are measured and entered into the prediction equation using the beta weights from the previous step. Predicted scores are then correlated with the essay scores assigned by human ratings, producing reliability estimates. Figure 1 shows a block diagram of the procedure.

The block diagram in Figure 1 is intended to show how essays are scored and validated using PEG. The shadowed blocks denote major sources of variation. In PEG, the major source of variation is the proxy variables used to compute the regression equation. The specific variables are not necessarily fixed across different sets of essays, and the final set of variables in the equation is determined empirically rather than theoretically. The lined boxes denote results of computations and are intended to show how results are used (e.g., "Validation essay scores" are correlated with "Predicted essay scores" to obtain a reliability measure). The remaining plain blocks denote intermediate steps in the scoring process.

Early results. The initial test of PEG used 276 essays written by students in Grades 8 to 12 (Page, 1966). Each essay was scored by at least four independent raters using a holistic rubric of overall essay quality. The rater scores were then summed to form the criterion score and proxy variables measured for all essays. A regression analysis was then done on a randomly drawn sample of half the essays (n = 138). The criterion score was entered as the dependent variable and the proxy variables as the independent variables. This sample was used to "train" the system. The next step was to compute the predicted score for the remaining essays. The proxy variables were entered into the prediction equation using the beta weights from the training phase. The multiple R for the predicted score was .71. The shrinkage was .65, and the number of proxy variables was 31. The predicted score correlated in the .50 range with the human raters, which was comparable to the correlation among human raters.

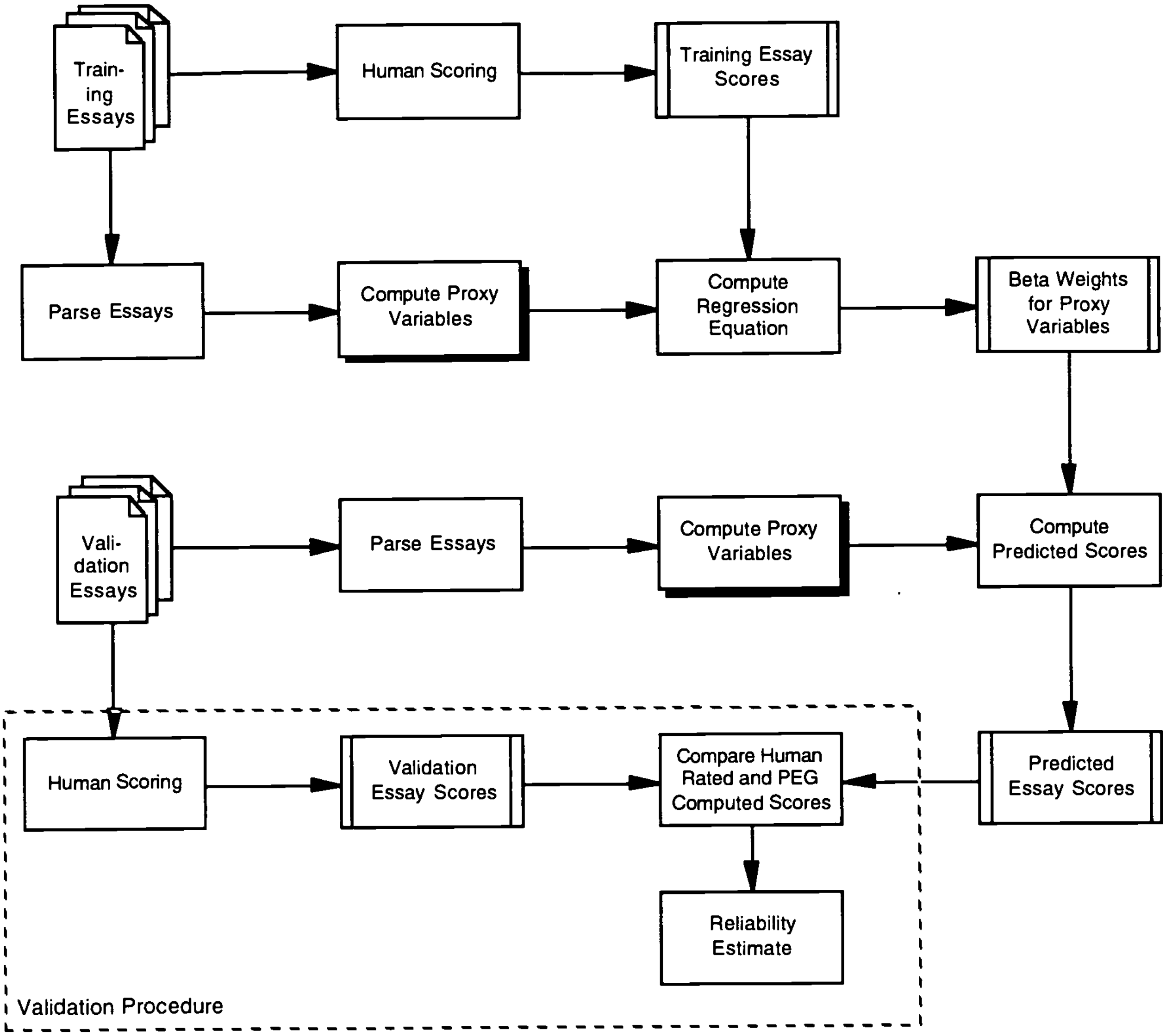

Figure 1. PEG scoring block diagram. Shadowed blocks indicate major sources of variation. Blocks with bars denote results of computations.

In another study used to test PEG (Page, 1968), 256 essays were scored by eight raters across five traits: ideas, organization, style, mechanics, and creativity. Using a similar procedure to Page (1966) to develop the beta weights, the predicted score for each of the traits ranged from .62 to .72, the shrunken multiple *R* from .55 to .69, and the corrected multiple *R* that takes into account the variance in the human ratings ranged from .64 to .78. Note that the correlations among human raters ranged from .72 to .85.

Current work. Page's early attempts at essay scoring point to a promising approach to scoring essays. More recently, PEG has shown even more impressive results. Using data from National Assessment of Educational Performance (NAEP) essays (Page, 1994), Page examined how PEG compared to trained raters. The topic of the essay was whether a city government should spend its recreation money on upgrading an abandoned railroad track or on converting an old warehouse for other uses. The essay had no correct answer. 599 essays were scored by 6 raters using a holistic rubric. The raters correlated among themselves between .46 and .67, with an average *r* of .56. To develop the prediction equation, two thirds of the essays were randomly chosen. Twenty proxy variables were used in the regression equation. The remaining essays (n = 189) were used to cross-validate the performance of PEG. This procedure was carried out three times, and the correlation between PEG and human raters for each trial was .86, .86, and .85. Another analysis examined how PEG scores correlated with individual raters and groups of raters. Using the same procedure but with 26 proxy variables, the predicted scores correlated with individual raters from .54 to .74, with an average *r* of .66. The range of correlations among human raters ranged from .39 to .65, with an average *r* of .55.

For pairs of raters, PEG scores correlated with each pair from .72 to .80, with a mean *r* of .75. Correlations among human raters ranged from .58 to .72, with a mean *r* of .66. For triplets (i.e., three human raters), the correlation between PEG and the raters was .81 and .79. In all cases PEG predicted individual human scores better than the human raters predicted each other.

In another set of studies, Page and Petersen (1995) conducted similar analyses and obtained similar results using Praxis essays from ETS. In this case, from a sample of 1314 essays, 1014 were used to train the PEG system. Each essay had at least two ratings. The remaining 300 essays were used to compare the performance of PEG and human raters. These 300 essays were scored by six raters, and these ratings were compared against the predicted score generated by PEG. Two examples of Praxis prompts are given below (Petersen, 1997):

Example 1. Some people argue that giving grades to students puts too much emphasis on competition and not enough emphasis on learning for its own sake. Others argue that without a precise grading system, students would not work as hard to excel in their studies because they would not have a standard against which to measure their performance. Should letter grading systems be replaced with pass/fail grading systems? Support your point of view with specific reasons and/or examples from your own experience, observations, or reading.

Example 2. Read the two opinions stated below:

Speaker 1: State-sponsored lotteries are a painless way to raise revenues while offering people a chance to strike it rich.

Speaker 2: Lotteries prey on the hopes of people who can least afford to buy tickets, and they encourage compulsive gambling.

If your state were considering introducing a lottery for the first time, how would you vote? Explain your position using specific reasons and/or examples from your own experience, observations, or reading.

The PEG scores correlated with individual raters from .72 to .78, with an average *r* of .74. Correlations among human raters ranged from .55 to .75, with an average *r* of .65. For pairs of raters, PEG correlated with each pair from .80 to .83, with an average *r* of .82. The correlations ranged from .77 to .82, with an average *r* of .78. For triplets, the correlation between PEG and the raters was .85 and .84. Correlations among the two sets of human raters was .85. Thus, as in the NAEP study, PEG predicted individual human raters' scores better than human raters predicted each other.

Strengths of PEG

Page's PEG system has several attractive features. The most appealing is that the predicted scores correlate as high as scores assigned by human raters and often higher. Other features of PEG are that the system can be used with widely available technology, is well understood methodologically, and will likely be improved as it incorporates more sophisticated kinds of measures. These issues are discussed next.

Predicted scores comparable to human raters. This is the most compelling aspect of PEG. Even the first-generation PEG system performed well (Page, 1966). More recent results based on NAEP and Praxis/ETS essays demonstrate even better performance (Page, 1994; Page & Petersen, 1995; Petersen, 1997). There is little doubt that PEG can predict scores on essays as well as, and often better than, humans.

Computationally tractable. Another attractive feature of PEG is that the system employs multiple regression to develop regression weights and then uses these weights in a prediction equation to compute scores. Neither the training stage nor the prediction stage seems computationally intractable. Although the processing time increases as the number of essays increases, computing the weights of the regression phase during the training phase is likely to be the most computationally intensive part of the system. Computing predicted scores is trivial, and parsing each essay to generate measures of each proxy variable is likely to take the longest time. From a practical standpoint, there is nothing about PEG that suggests it could not perform adequately on a standard desktop personal computer ($3000 range in 1997 prices).

Scoring methodologically is straightforward. The method used to score essays is procedurally very straightforward. The procedure uses a two-stage process, a training stage and a scoring stage. During the training stage, a sample of essays is drawn from the full set of essays. The criteria for selecting the training sample size are never provided by Page, but reported training sample sizes ranged from 50% (Page, 1968) to 77% (Page & Petersen, 1995). During the training stage the set of proxy variables must be selected as well. The optimal set of proxy variables is never suggested. Presumably, for a given set of training essays, a full set is initially entered into the regression equation and proxy variables that do not contribute to predictive power are dropped from the final regression equation. The dependent variable is the score assigned to the essays by a human rater. If there is more than one rater, then some combination of the scores is used (e.g., the mean or sum). The output of the training stage is a set of beta weights for the proxy variables from the regression equation.

During the scoring stage, proxy variables are measured for each unscored essay. The proxy variables are entered into the prediction equation, and a score is computed using the beta weights from the training phase. This value is the predicted essay score.

PEG is likely to improve. It is unlikely that the PEG system has reached maturity. Because of the simplicity of the proxy variables at this point, we think it is reasonable to expect that as more direct measures of essay style and content constructs are incorporated into PEG, the better PEG will be able to predict essay scores. Also, the incorporation of more direct measures of writing style and content can only improve the face validity of PEG.

Weaknesses of PEG

Although the performance of PEG remains impressive, there are limitations. First, the system suffers from a lingering question of how PEG can predict essay scores by simply examining the surface features of the essay, without considering meaning or content. Second, the system must be recalibrated each time for a new dataset. Third, the system can only provide scores for an essay that are relative to other essays. Finally, PEG is closed. The only published list of variables is for an early version of PEG (Page, 1968). These issues are elaborated next.

Construct objections. One criticism of the PEG system is that it does not take into account the semantics of essays. The meaning and content of the essays are ignored, and instead, only the surface features of the essay are considered. This suggests a relationship between score and essay that is nothing more than a statistical artifact. That is, the proxy variables are only statistically related to the true score of the essay.

Needs to be trained. The PEG system probably needs to be trained for each essay set used. This is inherent in the regression approach, which needs existing data to develop beta weights. Also, this means that the set of proxy variables that enter into the regression equation is not necessarily fixed and may change across different data sets.

Relative scoring method. Another inherent limitation of PEG is that it uses a relative scoring method. Scores can only be predicted if there exist other essays from the same (essay) population that can be used to train the PEG system. There is no way to determine the quality of an essay relative to an absolute or fixed criterion. For example, if students write essays based on source materials, there is no way for the PEG system to determine the extent to which the essays cover the content in the source materials. The PEG system can provide only a score based on the performance of other essays drawn from the same essay population.

Closed system. Finally, PEG is a closed system. Aside from Page (1966), there has been no publication of the detailed list of proxy variables used in the regression equations. There is no information about which proxy variables are used, much less the relative effectiveness of different proxy variables.

In the next section we review the latent semantic analysis system. Originally designed for information retrieval purposes, its recent application has been to text processing and the measurement of knowledge (Landauer & Dumais, 1997).

Latent Semantic Analyses (LSA)

Latent semantic analysis (LSA) is a technique originally developed for information retrieval purposes (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Latent semantic analysis represents information as a matrix that explicitly relates words and documents. Using this matrix formulation, LSA can compare a set of words (i.e., queries) against this matrix to determine the closest matching document. An assumption of LSA is that there exists some latent semantic structure in the data (i.e., meaning), and LSA attempts to capture that structure in the matrix representation. Besides traditional information retrieval uses, LSA has been used for cross-language information retrieval (Berry & Young, 1995), information filtering (Foltz & Dumais, 1992), classifying protein sequences (Wu, Berry, Shivakumar, & McLarty, 1995), and text analysis and essay scoring (Foltz, 1996; Landauer et al., 1997). The use of LSA or information retrieval techniques in general to score essays is a novel application. In this case, the specific information retrieved is not of interest; rather, what is important is the extent to which an essay can be matched against other essays, which may already be scored, or against a source document that the essay is based on.

Vector-space model. LSA is a variant of the *vector-space model* of information retrieval (Salton, 1991; Salton & McGill, 1983; Wong, Ziarko, Raghavan, & Wong, 1987). In a vector-space model, the information space is represented as a term-by-document matrix (see Figure 2). Note that "term" and "document" are used abstractly. A term could be a word, phrase, sentence, or any other unit of information. Similarly, document could be a sentence, paragraph, essay, or any other unit of information. The matrix arrangement simply sets up a mapping between a unit of information and its associated container.

	doc_1	doc_2	doc_3	•••	doc_n
t_1	w_{11}	w_{12}	w_{13}	•••	w_{1n}
t_2	w_{21}	w_{22}	w_{23}	•••	w_{2n}
t_3	w_{31}	w_{32}	w_{33}	•••	w_{3n}
⋮	⋮	⋮	⋮		⋮
t_m	w_{m1}	w_{m2}	w_{m3}	•••	w_{mn}

Figure 2. Term-by-document matrix. t represent terms, doc represent documents. Individual cells in the matrix contain term weights.

In the vector-space model, each row in the matrix represents a term and each column represents a document. Thus, a document is represented by a vector of terms, and if a cell within the vector contains a value greater than 0.0, then that term is present in the document. The particular value in the cell is determined by the specific term-weight applied (Lochbaum & Streeter, 1989; Salton & Buckley, 1988; Salton & McGill, 1983; Singhal, Salton, Mitra, & Buckley, 1996). The simplest term-weight is binary, where the cell is assigned a value of 1 or 0, depending on whether the term is present or absent in the document. Other term-weights are based on frequency of term occurrences. In general, the value of a term-weight reflects the term's importance relative to (a) the entire document set, (b) the document itself, or (c) a combination of both. Ideally, the weighting scheme would assign higher weights to terms that are more important to the content representation and lower weights to less important terms.

To retrieve documents, a query vector is set up in the same form as a document vector (i.e., a column in the term-by-document matrix). A query vector has the form shown in Figure 3. Cells in the query vector represent occurrences of a term. In Figure 3, t_1 ... t_m represent the same terms as those in the term-by-document matrix. The query weights, $qw_1 \dots qw_m$, may be binary (presence or absence of term) or may reflect the relative importance of each term within the query itself.

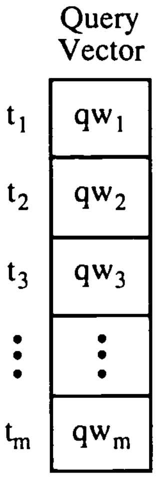

Figure 3. Query vector.

Once the query vector is set, a comparison is made between the query vector and each document in the term-by-document matrix, as shown in Figure 4. The resultant vector contains the similarity score between the query vector and each document. The higher the similarity score, the closer the match between the query and document. Various similarity measures exist, and the specific choice is driven empirically rather than theoretically (Jones & Furnas, 1987; Salton, 1991). A commonly used measure is the cosine coefficient, which is used in LSA. The cosine coefficient is the dot product between the query vector and each document vector in the matrix.

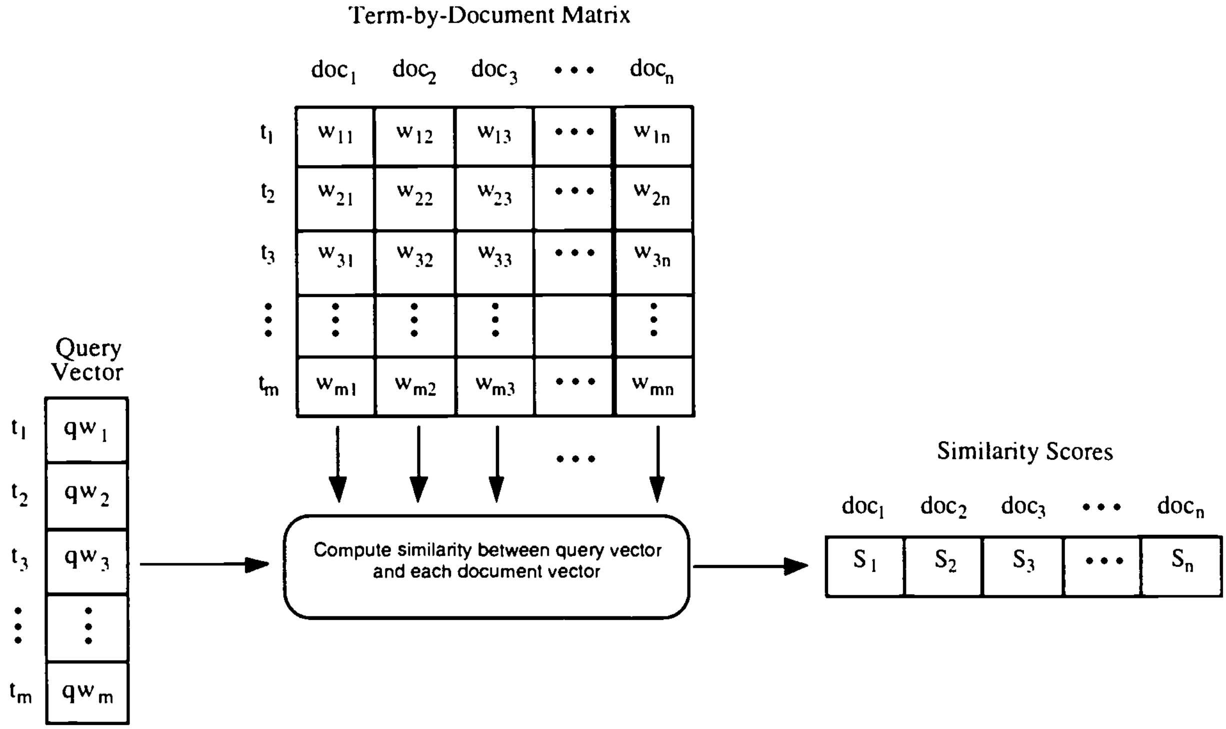
Term-by-Document Matrix
doc_1 doc_2 doc_3 ••• doc_n
t_1 w_{11} w_{12} w_{13} ••• w_{1n}
t_2 w_{21} w_{22} w_{23} ••• w_{2n}
t_3 w_{31} w_{32} w_{33} ••• w_{3n}
t_m w_{m1} w_{m2} w_{m3} ••• w_{mn}
Query Vector
t_1 qw_1
t_2 qw_2
t_3 qw_3
t_m qw_m
Compute similarity between query vector and each document vector
Similarity Scores
doc_1 doc_2 doc_3 ••• doc_n
S_1 S_2 S_3 ••• S_n

Figure 4. Computing similarity scores between the term-by-document matrix and query vector.

Because the matrix operations return a vector of scores, these scores can be used in a variety of ways. For example, in information retrieval applications, the scores are sorted and the top n scores returned to the user as representing the best matches. Or, the documents that are above a given threshold are returned to the user. In an essay scoring application, the same technique could be used to score documents. For example, each unscored essay could be considered a query vector, and the score of the essay could be the mean of the 10 most similar documents from the term-by-document matrix.

Given the basic vector-space model, LSA extends this model in a fundamental way, which has retrieval and computational implications. LSA employs *singular-value decomposition* to get an estimate of the original matrix, thereby (a) allowing for the retrieval of documents that may not contain any of the query terms, and (b) reducing computational complexity. In the remainder of this section we provide a minimal overview of the mathematics behind LSA. A complete discussion is provided in Deerwester et al. (1990).

LSA takes advantage of a fundamental property of matrices: Given an $m \times n$ matrix A, A has a singular value decomposition (Leon, 1994). This means that A can always be expressed as the product of three other matrices, $U\Sigma V^T$, where m = number of rows, n = number of columns, A = the term-by-document matrix, U = the term vectors, Σ = the singular values of A, and V = the document vectors. Note that Σ is a diagonal matrix containing singular values and is of rank K.

The essential point is that the original matrix A can be completely recovered by computing $U\Sigma V^T$ using the full rank of Σ (rank = K); however, by truncating Σ using a rank of k ($k < K$), an estimate of the original matrix A can be computed. In terms of data storage and computational requirements, this is a very desirable: The original matrix can be estimated from a smaller set of matrices. By setting k to values far smaller than the rank of A, computational time is reduced (as well as storage). This saving may be substantial given that matrix calculations are on the cube-order of the largest dimension of the matrix (Press, Teukolsky, Vetterling, & Flannery, 1992).

In LSA, Deerwester et al. (1990) suggest that truncating the singular value decomposition matrix reduces the amount of "noise" in the data while retaining only the most important "latent" structure of the information. By noise Deerwester et al. mean the variability in word usage across documents and queries. The reduction in noise helps with synonymy (many different ways to refer to the same object) and polysemy (many different meanings for one word). The singular value decomposition truncation also allows for the retrieval of similar documents *without the document necessarily containing any of the keywords of the query.*

Essay scoring with LSA. There have been several efforts to use LSA for text processing purposes (Foltz, 1996; Landauer et al., 1997), although none on the scale of PEG (Landauer & Dumais, 1997). We review two uses of LSA: identifying documents where information came from, and measuring essay quality. Figure 5 shows a block diagram of the general procedure.

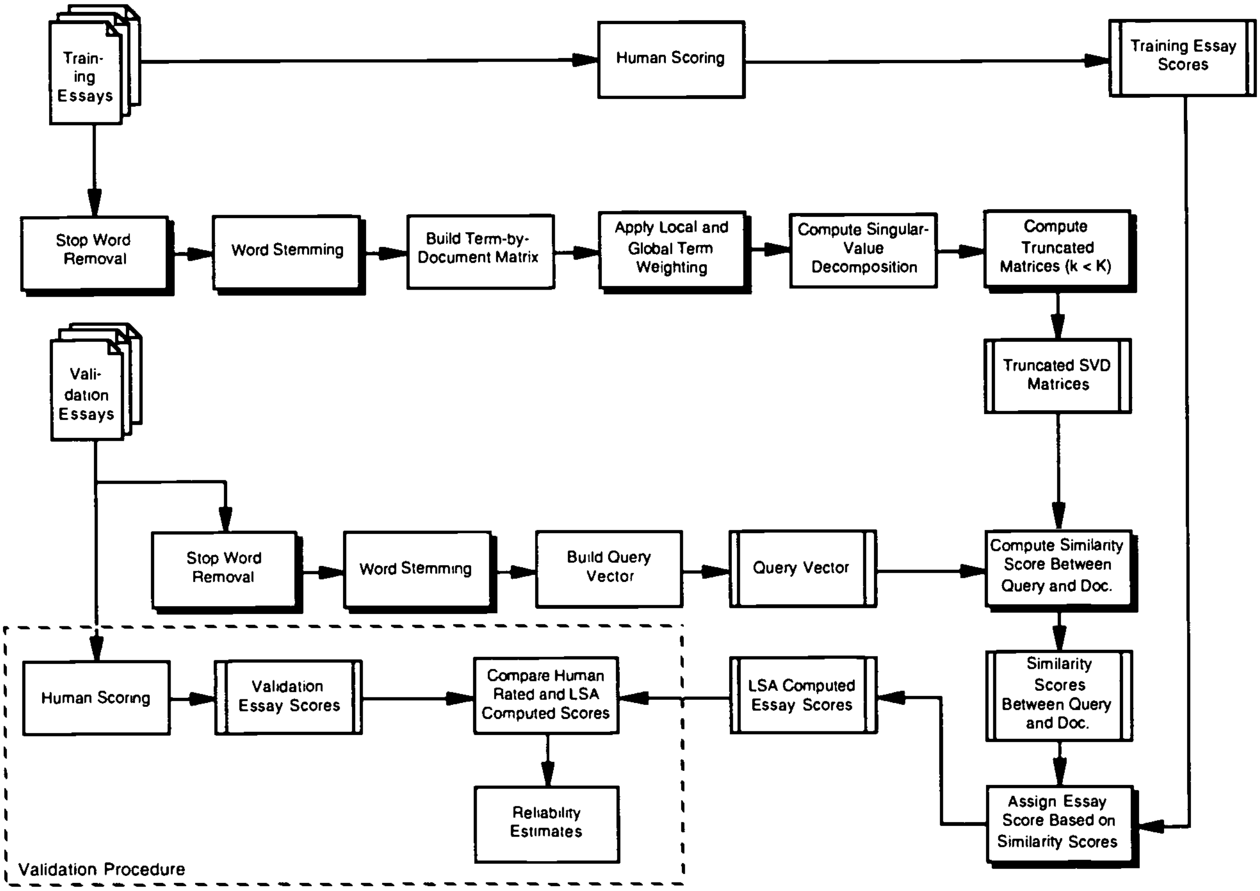
Training Essays
Human Scoring
Training Essay Scores
Stop Word Removal
Word Stemming
Build Term-by-Document Matrix
Apply Local and Global Term Weighting
Compute Singular-Value Decomposition
Compute Truncated Matrices (k < K)
Truncated SVD Matrices
Validation Essays
Stop Word Removal
Word Stemming
Build Query Vector
Query Vector
Compute Similarity Score Between Query and Doc.
Similarity Scores Between Query and Doc.
Assign Essay Score Based on Similarity Scores
LSA Computed Essay Scores
Human Scoring
Validation Essay Scores
Compare Human Rated and LSA Computed Scores
Reliability Estimates
Validation Procedure

Figure 5. LSA scoring block diagram. Shadowed blocks indicate sources of variation. Blocks with bars denote results of computations.

Identifying sources of information. This application of LSA sought to identify the document that was most likely the source of the essay content. Foltz, Britt, and Perfetti (as cited in Foltz, 1996) reported a reanalysis of data from an earlier study. Participants read 21 documents (6,097 words) covering the events that led to the construction of the Panama Canal. Participants were then required to write an essay answering the following question: "To what extent was the U.S. intervention in Panama justified?" The information space consisted of the original texts (6,097 words), supplemented with text from an encyclopedia (approximately 4,800 words) and excerpts from two books (approximately 17,000 words). The final singular-value decomposition matrix was 100 dimensions consisting of 607 documents (columns) and 4,829 words (rows).

A comparison was done between LSA and human raters on identifying where the information in the essay came from (Foltz et al., as cited in Foltz, 1996). Two raters examined all sentences for each individual. Each rater was asked to specify one or more of the 21 documents that were most closely tied to the sentence. Agreement between raters was computed liberally. If there was a match between any of the documents identified by either rater, then that was considered an agreement. Using this criterion, the human raters' agreement rate was 63%, and the average number of documents cited as being the source of a sentence was 2.1. LSA was then used to predict the document sources. The top 2 documents were picked and compared against each rater using the same agreement criterion. The overlap between LSA-selected documents and each human rater was 56% and 49%.

Measuring essay quality. To measure the quality of essays, Foltz et al. (as cited in Foltz, 1996) compared human-assigned scores with LSA-derived scores. Four history graduate students were used as expert graders. For each essay, they assigned a numeric score on two measures: (a) the information contained in the essay, and (b) the quality of the information. Each rater also read through all 21 texts and selected the 10 most important sentences he or she thought would be most useful in writing the essay.

Foltz et al. (as cited in Foltz, 1996) used two measures to capture different aspects of the essay quality. The first measure was a sentence-by-sentence comparison between each sentence in an essay and each sentence in the original text. The average of the similarity scores (i.e., the cosine measure returned by LSA) was used as a measure of the amount of recall. The second measure compared each sentence in the essay with the 10 sentences selected by the expert grader. For this measure, all sentences were compared against the 10 expert selected sentences. The score for each sentence was the highest similarity score between the sentence and all 10 sentences. The score for the essay was based on the mean of all sentence scores.

Foltz et al. (as cited in Foltz, 1996) then computed correlations among human raters, which ranged from .38 to .77. Correlations between human raters and the LSA scores on the text recall measure ranged from .12 to .55, with correlations between 2 of the 4 raters being statistically significant. Correlations between the human raters and LSA for the expert sentence measure ranged from .24 and .63, and were statistically significant for 3 of the 4 raters.

In a recent study that used a different method of scoring essays, Landauer et al. (1997) examined correlations between human raters and LSA. In the first experiment, 94 students were asked to write a 250-word essay on the function and anatomy of the heart. In addition, a 40-point short-answer test was administered and was used as an external criterion for both human raters and LSA. All raters were trained and graded each essay using a 5-point scale.

The LSA system was trained using 27 articles from an encyclopedia, resulting in a truncated matrix of 830 documents (sentence level) and 3,034 terms with *k* set to 94. Each sentence was considered a document (i.e., a column in the matrix). Two methods were used to compute scores. The first method involved assigning a score to an essay based on the average (human-rater assigned) score of the 10 most similar essays. The second method involved computing similarity scores between each essay and a section on the heart from a college biology textbook. The actual LSA-assigned score was based on two factors: (a) the similarity scores computed from either method, and (b) the vector length of the essay. The vector length is a function of the number of terms, which Landauer et al. (1997) interpret as being a measure of domain content.

Correlations were then computed for scores among human raters, and between LSA and each human rater. This was done for both methods. For the first method (scores based on other similar essays) correlation among human raters was .77, and between LSA and each human rater .68 and .77. The correlation between LSA and the average rater score was .77. The correlations between the 40-point short-answer (external criterion) test and the average rater score and LSA were .70 and .81 respectively. For the second method (scores based on textbook section), the correlation between LSA scores and each rater was .64 and .71, and between LSA and the average rater score .72. The correlation between the LSA score and the 40-point short-answer (external criterion) test was .77.

In a second study with different content, Landauer et al. (1997) used the first scoring method to compute essay scores. Two hundred seventy-three psychology students were required to write essays on one of three topics. Two raters scored each essay. One rater was either the instructor or a graduate teaching assistant. The other rater was one of two advanced undergraduate psychology major teaching assistants. The LSA system was trained on the textbook used in the course. The source material consisted of 4,904 paragraphs and 19,153 unique terms with a *k* of 1500.

For all essays, the correlation among human readers was .65 and the correlation between LSA and the average reader score was .64. For the first essay, the correlation among human raters was .19, and between LSA and the average reader score .61. For the second, .75 and .60, and for the third essay, .68 and .71.

Strengths of Latent Semantic Analysis

Latent semantic analysis has several attractive features that look promising with respect to automated scoring of essays. In particular, LSA seems to be comparable to human scoring, offers flexible assessment options, and employs matrix operations that are well understood and available. Each of these strengths is discussed next.

Scoring performance similar to human raters. As the results of Foltz et al. (cited in Foltz, 1996) and Landauer et al. (1997) show, LSA can predict scores that correlate well with human assigned scores. We view these findings as promising but preliminary.

Relative and absolute scoring. One of the nicest features of LSA is that, for scoring purposes, comparisons can be made relative to other essays, or against a known reference (e.g., a textbook). When a relative scoring method is used, a target essay is compared against other essays, and the score assigned to the target essay is based on the scores of the *n* most similar essays. When an absolute scoring method is used, a target essay is compared against a fixed source such as an expert written essay, source materials, or another criterion.

Relatively open system. The two major subsystems of the LSA methodology are in the public domain and readily available. The vector-space representation has existed in information retrieval systems since the late 1960s (Salton, 1971). The singular-value decomposition procedure is a property of matrices, has been used extensively in engineering applications (Press, 1992), and exists in statistical packages (e.g., SPSS, 1990) as well as in source code form (e.g., Press et al., 1992). What is novel and proprietary about LSA is the way singular-value decomposition is being used specifically for information retrieval purposes (U.S. Patent No. 4,839,853, June 13, 1989).

Weaknesses of Latent Semantic Analysis

Although LSA has some attractive features, the entire approach is computationally expensive. LSA requires large amounts of information and large amounts of computational power. The use of matrices and vectors to represent the data and to determine similarities requires a tremendous amount of computation. However, with the advance of computer power (doubling over two years) this latter point will be less relevant in the future. In addition, LSA shares construct problems with PEG. Word order is omitted from LSA so it is difficult to conceptualize how "meaning" is taken into account. These issues are discussed in greater detail below.

Construct objections. One criticism of LSA is that the methodology does not take into account word order, which means that every possible combination of words in a given sentence is equivalent. This arises from the matrix representation of the information. However, Landauer and Dumais (1997) assert that the latent structure that emerges from the singular-value decomposition scaling procedure (i.e., the truncated singular-value decomposition matrix) represents an underlying structure of the information in the document, and it is this latent structure that represents meaning. Further, the scores between human ratings and LSA-generated scores are comparable, suggesting that LSA is sensitive to the information in the documents.

Large number of examples needed to calibrate system. Another weakness of LSA is that the techniques require many examples of word usage or word co-occurrences. There must be a large corpus of text available to train the system on.

For example, Foltz et al. (as cited in Foltz, 1996) supplemented the system with material from encyclopedia articles and book excerpts. In fact, the supplementary text comprised 72% of the materials (in words) used to train the system. Similarly, Landauer et al. (1997) used large numbers of articles to train the LSA system. In the first experiment, 27 articles from an encyclopedia were used, and in the second experiment, an entire textbook was used.

Computationally expensive. Finally, LSA is computationally expensive. Matrix computations are inherently CPU intensive, with the number of operations proportional to the third power of the matrix size (Press et al., 1992). Dumais (1994) reports that for huge matrices (60,000 × 80,000), computing the singular-value decomposition took 18 to 20 hours of CPU time and 250MB of RAM on a Unix workstation. Although this provides an estimate for huge matrices, our local feasibility tests using much smaller matrices (approximate 100 documents × 2,000 terms) and SPSS on a Macintosh Centris 650 with 24MB of memory ran over 24 hours without completion. Using a similar-sized matrix on the IBM RS/6000 clustered UNIX workstations, SPSS took about 10 hours of CPU time with 128MB of RAM. The most computationally expensive part of LSA is calculating the singular-value decomposition and associated matrices, which must be done every time the original term-by-document matrix is altered (e.g., by adding new documents or recomputing the term weights). In addition, each query must be multiplied against each document vector to compute the similarity ratings. This time is a function of the number of documents in the term-by-document matrix.

Summary of Methodologies

The PEG system and LSA offer two different approaches to automated scoring of essays. PEG uses a regression model that predicts scores based on a set of proxy variables. These proxy variables are assumed to reflect the intrinsic constructs of an essay (e.g., style) as reflected in surface features of the text. For example, the fourth root of the length of the essay, use of uncommon words, and use of prepositions are all good predictors of essay score (Page & Petersen, 1995). Similarly, use of uncommon words and use of prepositions were shown to be good predictors of essay quality (Page, 1966). In contrast, LSA bases its scores on word co-occurrences in essays. A term-by-document matrix is created that explicitly maps the relationship between a word and its document. Then a scaling procedure, called singular-value decomposition, is carried out on the matrix. This scaling procedure produces a set of smaller matrices that can be used to estimate the original matrix, but with far fewer dimensions. Queries are carried out by computing the similarity between each query and each document in the truncated matrix. Geometrically, each document can be interpreted as a vector in *k*-dimensional space. A query, represented as a vector, is mapped onto this space and the cosine between the query vector and each document vector is computed. The smaller the angle between the two vectors, the higher the similarity between the query and document.

Performance of PEG and LSA, as essay scoring technologies, is quite impressive. Both methods produce scores that correlate highly with human raters, with PEG scores generally correlating better with human raters than the human raters among themselves. LSA performance is just the opposite, with LSA scores generally correlating as well or slightly lower with human raters than the human raters among themselves. Regardless of system, essay scores predicted by either tend to be as good as human judgments. The PEG system, we believe, has ample room for performance improvement and is constrained by software technology. As the software becomes available to extract more sophisticated forms of information from the essay, the PEG system will be able to incorporate more direct measures of essay constructs. With regard to LSA, the research program is in its infancy and is likely to produce better essay scoring performance as well as interesting text processing applications.

However, both methods have some drawbacks. The PEG system is closed. Little is published about specific variables that comprise the regression equations. Also, the system is limited in two ways. First, PEG is designed as an essay scoring system and nothing else. The research has focused exclusively on how to predict scores on a target essay given measures on a variety of surface characteristics. This has no applicability to other text processing applications (e.g., categorizing single sentences or other text materials that are not part of a larger body of text). The second limitation is that the essay scoring system is exclusively relative. Scores assigned to an essay are based on the characteristics of other essays drawn from the same essay population. Scores cannot be predicted from the comparison of a target essay and a fixed source (e.g., source materials or expert-written essays).

The LSA system has several drawbacks as well. The approach is computationally expensive and requires RAM on the order of hundreds of megabytes, unless optimization techniques are used (e.g., Kolda, 1997). The second drawback is that LSA needs supplementary source material for numerous examples of appropriate word co-occurrences.

Error Analysis

This section discusses potential sources of errors. Sources of variation exist at all stages of scoring. Table 2 summarizes sources of variation for PEG and LSA that we think may have a substantial impact on scoring performance.

Table 2

Sources of Variation for PEG and LSA

Variable	Impact on PEG	Impact on LSA
Human rating of training essays	Substantial. The regression model is based on the scores of the training essays, and the scores are directly associated with the predictor variables during training. The error will be incorporated in the regression equation; thus, the predicted scores will reflect this error. Using more raters to score the documents will reduce this error.	Same as PEG.
Essay characteristics	**Content.** Not handled by PEG. However, implicitly handled by the human-scored essays used to train system.	**Content.** Substantial. LSA relies on word co-occurrences within and between documents.
	Surface characteristics. Substantial effect on performance. Surface characteristics form independent variables in regression equation.	**Surface characteristics.** No impact. LSA looks only at words, not punctuation or other surface elements.
	Essay length. Substantial effect on performance. The 4th root of length of the essay is a good predictor of essay score.	**Essay length.** Substantial. If essay length is not adjusted for, longer essays will have higher term-weights. Essay length can be adjusted by term-weighting that takes into to account document length. Also, the term-by-document matrix structure can be used to take into account document length (i.e., whether a document is a sentence, paragraph, or essay).

Table 2 (continued)

Variable	Impact on PEG	Impact on LSA
	Word usage. Substantial. For example, uncommon words and prepositions show up as good predictors of essay performance.	**Word usage.** Substantial impact. Essays with a large number of words that overlap with the criterion essays will be have a higher similarity score than essays with a lower overlap. Very common words typically omitted from term-by-document matrix.
	Punctuation. Substantial. Use of different punctuation comprises several proxy measures.	**Punctuation.** No impact. LSA does not consider punctuation.
Document pre-processing	**Stemming**. Unknown, but probably substantial. Removal of suffixes alters word length and possibly distinction between common and uncommon words.	**Stemming**. Unknown, but Lochbaum and Streeter (1989) and Hull (1996) suggest that LSA performance will benefit from stemming.
	Removal of stop-words. Unknown, but probably substantial. Prepositions are good predictors of essay score, but they are typically considered stop-words.	**Removal of stop-words.** Substantial. Removal of common words is important because these words contribute nothing to distinguishing more and less important terms. Removal of stop words reduces matrix size and thus improves overall performance.
	Term-weighting. Not applicable.	**Term-weighting**. Substantial. Different term-weightings have markedly different effects on LSA results (Lochbaum & Streeter, 1989; Salton & Buckley, 1988; Singhal et al., 1996).
System training	**Number of training documents.** Substantial. There is probably a threshold below which the standard error of the regression equation is too large for reasonable predictions. However, the value of this number is unknown.	**Number of training documents.** Substantial. LSA needs many different examples of appropriate word co-occurrences. This can only be done by training LSA on a large number of documents.
Classification parameters	**Similarity score.** Not applicable.	**Similarity score**. Substantial. Different similarity scores have resulted in wide swings in what is considered "similar." However, the cosine measure has been used in a variety of systems, including LSA and SMART (Deerwester et al., 1990; Jones & Furnas, 1987; Salton, 1971)

Other Potential Applications

In this section we discuss the more general problem of text classification and consider possible applications. These applications, unless otherwise noted, assume an LSA-based system. LSA offers a general analysis technique that appears suitable for measuring a variety of text-based information.

Short-Answer Responses

Short-answer responses typically consist of a brief response of a few sentences. One application is to use these responses as a quick measure of prior knowledge (e.g., Baker, Aschbacher, Niemi, & Sato, 1992). Some characteristics of short-answer responses are (a) few sentences, (b) very specific prompt asking for specific information, (c) sentences not fully formed. Our prediction is that PEG would not work well with this kind of data, primarily because there would not be enough variability in the proxy variables given the limited text and the likely existence of ill-formed sentences. We think that for short-answer responses, LSA would perform better. The same technique used for scoring full essays could be employed, specifically, the use of source text as the criterion. What remains an empirical question is how well LSA would perform relative to human raters.

Typed Responses in a Collaborative Environment

This application would categorize the communication (messages) between users in a computer-based, networked environment. The purpose of categorizing such communication is to obtain measures of collaboration and teamwork, which would reduce the amount of labor involved in manually categorizing the messages. Our initial approach would be to use as the source text a bank of existing messages developed for the measurement of teamwork skills (Chung, O'Neil, Herl, & Dennis, 1997; O'Neil, Chung, & Brown, 1995, 1997). This set of messages, in addition to scored protocols of typed messages, would be used to train the system. Categorizing would occur by assigning the target protocol the category of the most similar protocol or set of protocols. Other sources of online data would be bulletin boards, email, and real-time chat.

Verbal Protocol Data

As with the typed messages, this approach would categorize think-aloud protocols of subjects engaged in a particular task. The approach would be similar to the measurement of typed messages: A set of scored protocols would be used to train the system. Scoring would occur by assigning the target protocol the category of the most similar protocol or set of protocols.

Scoring of Information Sources

Another application of LSA is to score information sources (e.g., Web pages) that subjects use while carrying out some task (Schacter et al., 1997). The system would be trained on accepted standards such as encyclopedia articles or textbooks. Each information source could then be scored against this reference. This process would produce a measure of relatedness or relevance of the information sources.

Measures of Free Recall

Another application of LSA is to use the approach of Foltz et al. (as cited in Foltz, 1996) with LSA to yield measures of free recall. In their approach, Foltz et al. performed a sentence-by-sentence comparison between the essay protocols and the source materials. The strength of the similarity between subjects' protocols and the original material would yield estimates of what in the material is most likely to be remembered, as well as what content subjects were spending their reading effort on.

Computer Trace Data

Another potential application of LSA would be to score online computer behavioral data. This approach conceptualizes a sequence of trace data as documents and particular events as terms. This application is highly speculative, but the premise is that behavior reflects complex decision making and processing of information. If LSA is detecting latent knowledge and traits, as Landauer and Dumais (1997) claim, then the application of LSA to behavioral data would seem to be reasonable.

Proposed Investigations

In this section we outline a plan for investigations of automated processing of text-based data. If the question is the feasibility of automated scoring techniques for the singular purpose of scoring essays relative to other essays, with no interest in the analysis of other forms of text-based data, then there is little question that PEG is feasible and probably the best choice. PEG exists today and has demonstrated success across a variety of different essay formats. Ellis Page has a 30-year commitment to PEG and has probably developed a stable set of proxy variables for his regression model. The bottom line is that PEG can predict scores at least as well as humans, and oftentimes PEG is more reliable at predicting human scores than human raters are at predicting each other's scores.

However, if the question is the feasibility of automated scoring techniques for the purpose of analyzing a range of text-based data and potentially other forms of data, which includes but is not limited to essays, then we believe LSA to be the technology with greater potential. PEG is inherently limited by the use of a regression model, which means that a score can be predicted for a given essay, but only relative to the essays the system was trained on. There is no capability for comparing a given essay to an external criterion (e.g., a textbook). LSA on the other hand, uses a factor-analytic approach that can accommodate scoring essays relative to other essays, but within a much more general framework. Comparisons are possible not only between different kinds of textual data, but also between different-sized text chunks (e.g., single sentences, paragraphs, or long essays), and in its most general form, any kind of data, textual or not.

One shortcoming of LSA is that there has been little research on its use in text processing. We do not know whether there are limitations on the kinds of text it can use, the granularity of the text chunks used to form the term-by-document matrix or the queries, or the effects of the different kinds of terms weightings. Further, a very basic term-by-document approach is used, with term usually being associated with words. We know of no research that has examined how the granularity of the "term" (i.e., the use of phrases or other higher orders of meaning) affects performance. Thus, there are many factors that may affect the performance of LSA, but at this time, there is little basis from which to draw any firm conclusions about LSA.

Figure 6 is a block diagram, modified from Figure 5, of our suggested approach to the systematic study of LSA in text processing. Each shadowed box represents a factor that we believe impacts LSA performance. These factors form the basis for a set of small-scale studies that will investigate the performance impact of these variables on LSA. We explain each factor below.

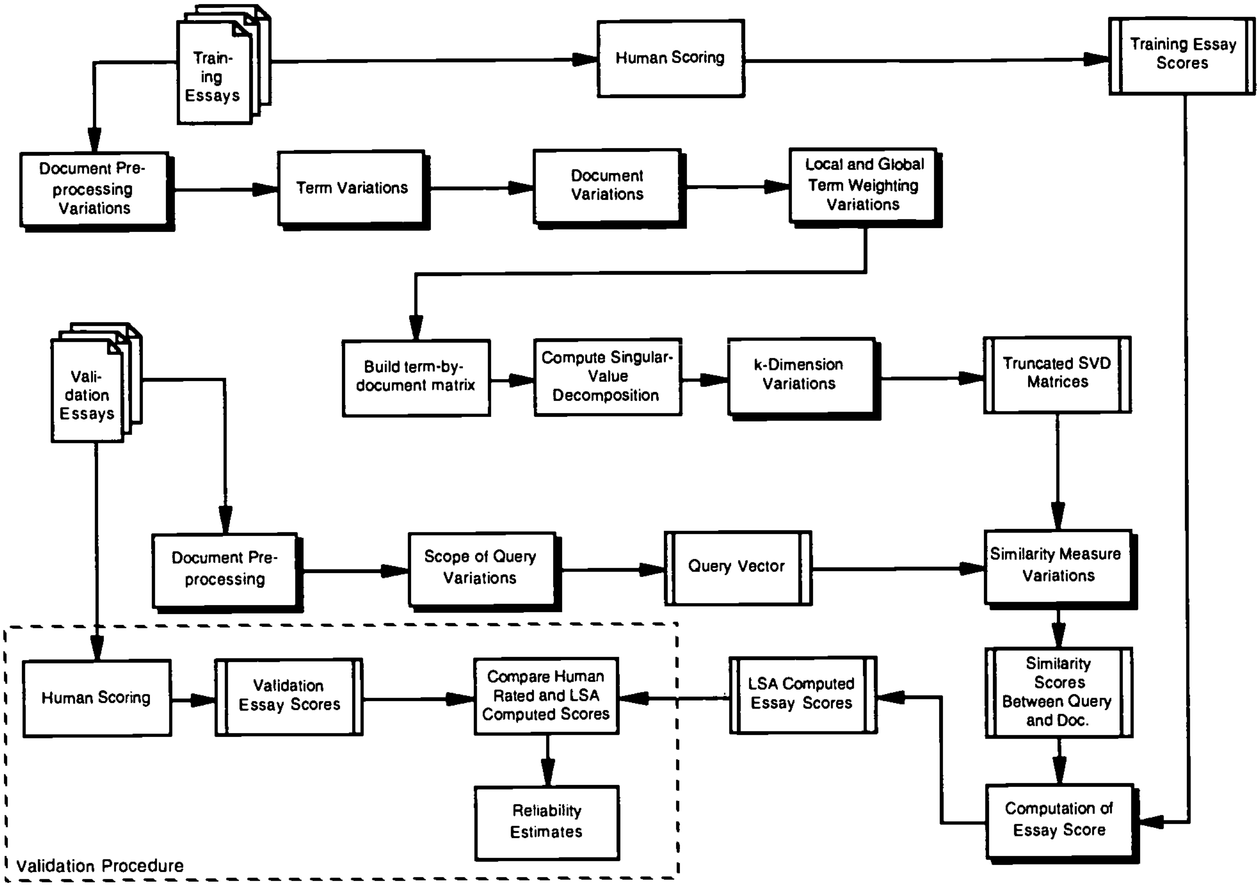
Train-ing Essays
Human Scoring
Training Essay Scores
Document Pre-processing Variations
Term Variations
Document Variations
Local and Global Term Weighting Variations
Vali-dation Essays
Build term-by-document matrix
Compute Singular-Value Decomposition
k-Dimension Variations
Truncated SVD Matrices
Document Pre-processing
Scope of Query Variations
Query Vector
Similarity Measure Variations
Human Scoring
Validation Essay Scores
Compare Human Rated and LSA Computed Scores
LSA Computed Essay Scores
Similarity Scores Between Query and Doc.
Reliability Estimates
Computation of Essay Score
Validation Procedure

Figure 6. Processing block diagram. Shadowed blocks indicate sources of variation. Blocks with bars denote results of computations.

LSA Performance Impact Variables

Document pre-processing. Existing approaches include word-stemming and removal of stop-words. Word stemming is the removal of suffixes, which reduces the number of unique words (Harmon, 1991). However, depending on the stemming algorithm, a word like "subjected" may be stemmed to "subject," a complete alteration of the original meaning. Removal of stop-words means dropping from the term-by-document matrix very common words. Stop-words are words that provide little information towards distinguishing one document from another. An example of a stop-word is "the."

Term variations. In vector-space models, term usually refers to a word. However, the definition of term could be broadened to include phrases, parts of speech, or other kinds of information. Perhaps, as in PEG, surface features could be included as well.

Another application is to include abstract categories as a "term." For example, Wendlandt and Driscoll (1991) found that the inclusion of category information improved retrieval performance substantially over normal keyword based systems. Wendlandt and Driscoll identified a set of categories (e.g., amount, location, time, purpose, source, duration, cause, result, purpose, space, goal) and words associated with each category (e.g., after–time; above–amount, location, time; before–location, time; any–amount; why–cause, purpose). This kind of information can be easily folded into LSA.

Document variations. Documents in a vector-space model can be of varying size, such as a sentence, paragraph, or essay.

Local and global term-weighting. Various term-weighting schemes have been developed to adjust for the importance of words relative to the document and relative to the entire collection.

***k*-Dimension variations.** For LSA, the best value of k is empirically determined. Deerwester et al. (1990) suggest the optimal value of k reduces the amount of "noise" or spurious word co-occurrences that exist in all collections while still being able to distinguish between documents.

Similarity measure variations. The similarity function usually used in vector-space and LSA implementations is the normalized cosine function. There is no theoretical basis for the use of cosine over other measures, other than it seems to work the best (Deerwester et al., 1990; Salton, 1989).

Computation of essay score. Given a set of pre-scored documents found to be similar to the unscored essay, how is the essay score computed? Landauer et al. (1997) used the average of the 10 most similar documents. Another possibility is scoring based on a cut-off value of the similarity score. In this case, the average of the scores of all documents above the cut-off value is the score assigned the unscored document.

Discussion

The use of automated scoring techniques for assessment systems raises many interesting possibilities for assessment. Essays are one of the most accepted forms of student assessment at all levels of education and have been incorporated in many of the standardized testing programs (e.g., the SAT and GRE). Issues of validity and reliability remain unchanged and must be addressed in order to fully deploy automated approaches to essay scoring. A paradox with the PEG and LSA systems is that neither system attempts to interpret the text, yet both approaches predict essay scores as well as human raters. Broad acceptance of these automated essay scoring systems will require a clear explanation of what variables are being measured and how those variables relate to a theoretical model of writing and cognition. A model is needed to explain the peculiarities of each approach. For example, PEG needs a model to explain why surface features alone (i.e., no content) are sufficient to predict essay quality. Likewise, for LSA, how can two sets of words with different word orders (even random order) yield equivalent essay scores? The alternative is to treat each system as a black-box, accept the ambiguity of the approach, and impose cross-validation checks throughout the scoring process to ensure reasonable performance.

Another issue raised is how general PEG and LSA are to other types of text processing tasks. Because of the widespread use of text in education and research, both as a source of information and as the product of student performance, a very desirable feature of the system is to handle multiple kinds of textual data. Landauer et al. (1997) and Foltz et al. (as cited in Foltz, 1996) have begun to examine the application of LSA to different processing tasks. However, because there has been no clear cognitive rationale for either approach, nor sufficient research on the effects of different system parameters on performance, any statements about the application of either approach to non-essay scoring tasks are speculative at best.

A final issue raised is that there still remains a need for human scoring. An assumption of this report and of both essay scoring approaches is that the human rating is the best estimate of the true essay score. Thus, there will always be a need for some portion of the documents to be scored using multiple trained raters. One exception is when the text comparison is done against a fixed source (e.g., an expert essay) or source materials.

The importance of the proposed investigation lies far more in the potential for practical spin-offs than in any theoretical contribution to writing, education, assessment, or cognition. We view automated scoring techniques as tools with which to help testing agencies, practitioners, and researchers handle text data.

From an assessment perspective, such a system would drastically reduce the cost of testing and increase the timeliness of reporting of results. From a teaching perspective, the availability of scoring tools would encourage teachers to assign more writing tasks than is done now. And from a research perspective, the availability of an automated scoring tool would make tenable investigations that would otherwise be rejected due to the costliness and labor intensive nature of scoring text data.

Broad acceptance of automated scoring techniques for essays and other textual materials will require that these issues be addressed. From a technological standpoint, a general approach to the classification of text documents is likely to be some time away. We believe that LSA and PEG represent first attempts to specifically address the essay scoring problem. Given their infancy, we are impressed by their current performance, and we view the direction of automated scoring techniques as extraordinarily promising.

References

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Bennett, R. E., & Bejar, I. I. (1997). *Validity and automated scoring: It's not only the scoring*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Berry, M. W., & Young, P. G. (1995). Using latent semantic indexing for mulilanguage information retrieval. *Computers and the Humanities, 29,* 413-429.

Carley, K. (1988). Formalizing the social expert's knowledge. *Sociological Methods and Research, 17,* 165-232.

Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science, 46,* 194-216.

Chen, H., Orwig, R., Hoopes, L., & Nunamaker, J. F. (1994). Automatic concept classification of text from electronic meetings. *Communications of the ACM, 37*(10), 56-75.

Chung, G. K. W. K., O'Neil, H. F., Jr., Herl, H. E., & Dennis, R. A. (1997, April). *Use of networked collaborative concept mapping to measure team processes and team outcomes*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Daigon, A. (1966). Computer grading of English composition. *English Journal, 55,* 46-52.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391-407.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason.* Cambridge, MA: MIT.

Dumais, S. T. (1994). Latent semantic indexing (LSI): TREC-3 Report. In D. K. Harmon (Ed.), *Text REtrieval Conference* (TREC-3). Gaithersburg, MD: U.S. Department of Commerce, Technology Administration, National Institute of Standards and Technology.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers, 28,* 197-202.

Foltz, P. W., & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM, 35*(12), 51-60.

Hardy. R. A. (1995). Examining the cost of performance assessment. *Applied Measurement in Education, 8,* 121-134.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42,* 7-25.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed description. *Journal of the American Society for Information Science, 47,* 70-84.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38,* 420-442.

Kaplan, R. M., & Bennett, R. E. (1994). *Using the free-response scoring tool to automatically score the formulating-hypotheses item* (ETS Tech. Rep. No. GRE 90-02b). Princeton, NJ: Educational Testing Service.

Kolda, T. G. (1997). *Limited-memory matrix methods with applications* (UMCP-CSD Tech. Rep. No. CS-TR-3806). College Park: University of Maryland, Department of Computer Science.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Newbury Park, CA: Sage.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). *How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans*. Paper presented at the 19th annual conference of the Cognitive Science Society, Palo Alto, CA.

Leon, S. J. (1994). *Linear algebra with applications* (4th ed.). New York: Macmillan.

Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan (Eds.), *Proceedings of the fourteenth annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 262-269). Baltimore, MD: AMC Press.

Lochbaum, K. E., & Streeter, L. A. (1989). Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing and Management, 25*, 665-676.

Martinez, M. E., & Bennett, R. E. (1992). A review of automatically scorable constructed-response item types for large-scale assessment. *Applied Measurement in Education, 5*, 151-169.

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. S. (1995). *Measurement of teamwork processes using computer simulation* (CSE Tech. Rep. No. 399). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Jr., Chung, G. K. W. K., & Brown, R. S. (1997). Use of networked simulations as a context to measure team competencies. In H. F. O'Neil, Jr. (Ed.), *Workforce readiness: Competencies and assessment* (pp. 411-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Orwig, R., Chen, H., & Nunamaker, J. F., Jr. (1997). A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science, 48*, 157-170.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238-243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*, 210-225.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*, 127-142.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*, 561-565.

Petersen, N. S. (1997). *Automated scoring of writing essays: Can such scores be valid?* Presentation at the annual meeting of the American Educational Research Association, Chicago, IL.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). New York: Cambridge University Press.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics, 61*, 241-254.

Roberts, C. W., & Popping, R. (1993). Computer-supported content analysis: Some recent developments. *Social Science Computing Review, 11*, 283-291.

Salton, G. (Ed.) (1971). *The SMART retrieval system. Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Salton, G. (1991). Developments in automatic text retrieval. *Science, 253,* 974-253.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24,* 513-523.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* San Francisco, CA: McGraw-Hill.

Schacter, J., Herl, H. E., Chung, G. K. W. K., O'Neil, H. F., Jr., Dennis, R. A., Lee, J. J. (1997, April). *Feasibility of a Web-based assessment of problem solving.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing and Management, 32,* 619-633.

SPSS reference guide [computer program manual]. (1990). Chicago, IL: SPSS.

Wendlandt, E. B., & Driscoll, J. R. (1991). Incorporating a semantic analysis into a document retrieval strategy. In A. Bookstein, Y. Chiaramella, G. Salton, & V. V. Raghavan (Eds.), *Proceedings of the fourteenth annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 270-279). Baltimore, MD: AMC Press

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1987). On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems, 12,* 299-321.

Wu, C., Berry, M., Shivakumar, S., & McLarty, J. (1995). Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning, 21,* 177-193.

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title:	Methodological Approaches to Online Scoring of Essays
Author(s):	Gregory K. W. K. Chung and Harold F. O'Neil, Jr.
Corporate Source: UCLA Center for the Study of Evaluation	Publication Date: December 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑ Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□ Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Kim Hurst	Printed Name/Position/Title: Kim Hurst/Admin. Assistant	
Organization/Address:	Telephone: 310-206-1532	FAX: 310-825-3883
	E-Mail Address: Kim@cse.ucla.edu	Date: 2/25/98